AfriSwitchLID: Multilabel Spoken Language Identification Models for Code-Switched African Conversations

Anonymous ACL submission

Abstract

Code-switching is a frequent characteristic of everyday speech across African multilingual conversations, yet existing spoken language identification (LID) models are designed and evaluated exclusively on monolingual speech. We present the first systematic evaluation of spoken CS-LID across 20 African languages and five natural African-English CS pairs. We convert multiclass to multilabel predictions by selecting the top-5 softmax scores, and surface two fundamental failure modes in existing monolingual approaches, (1) LID Degradation: Kinyarwanda LID accuracy, for example, drops 41 percentage points from monolingual to CS settings, and (2) Top-N miss: baseline models correctly identify both CS languages in the top-5 predictions for fewer than 10% of utterances across all evaluated pairs. Multilabel fine-tuning on synthetic CS data achieves 96.9% both-in-top-5 accuracy across four trained African-English pairs with negligible monolingual degradation, and partial zero-shot transfer to a held-out pair (Zulu-English: 35.7% vs. 1.8% baseline) demonstrates generalisation beyond trained language pairs.

1 Introduction

Multilingual speech is the norm rather than the exception across African communities, where speakers routinely alternate between languages within a single conversation or utterance; a phenomenon known as code-switching (CS) (Yılmaz et al., 2018; Li et al., 2023). Language identification (LID) is a critical prerequisite for downstream spoken language processing tasks including automatic speech recognition, machine translation, and spoken dialogue systems (Adebara et al., 2022). In multilingual African deployment contexts, where a single call center can serve a diverse speaker population spanning 10+ languages, a reliable LID system must handle not only clean monolingual speech but also the code-switched speech that characterizes everyday communication (Yılmaz et al., 2018). Where lingua francas exist (e.g. Nigerian Pidgin English), conversations could span more than 2 languages (e.g. English, Igbo, Pidgin) as speakers switch languages to better express ideas. Yet existing spoken LID models are designed and evaluated primarily on monolingual speech, leaving open the question of whether strong monolingual performance predicts reliable performance in code-switched settings (Adebara et al., 2022; Burchell et al., 2024).

Two broad approaches to CS-LID exist in current systems. The first assigns explicit bilingual labels to CS speech, training on paired CS corpora for specific language combinations (Jain and Bhat, 2014; Lin et al., 2014). The second applies models with monolingual label spaces directly to CS speech, relying on majority-language detection (Mager et al., 2019; Burchell et al., 2024). Both approaches have fundamental limitations: explicit CS labels do not generalise beyond their trained pairs, while monolingual label spaces cannot represent the simultaneous presence of two+ languages by design (Burchell et al., 2024). To our knowledge, no prior work has systematically compared these approaches on African language code-switched speech.

The contributions of this paper are as follows:

1. The first systematic multilabel evaluation of spoken CS-LID across 20 African languages and five natural African-English CS pairs

2. A taxonomy of two baseline CS-LID failure modes, (1) explicit label generalisation failure, and (2) softmax suppression failure, with empirical characterisation of each on natural African CS speech.

3. Multilabel fine-tuning on synthetic CS data, achieving 96.9% both-in-top-5 accuracy

across four African-English pairs with negligible monolingual identification degradation

4. Analysis of partial zero-shot transfer to a held-out language pair (e.g. Zulu-English), with characterisation of the conditions under which synthetic CS fine-tuning generalises beyond trained language pairs.

2 Related Work

African Language Speech Processing Datasets such as FLEURS (Conneau et al., 2023), CommonVoice (Ardila et al., 2020), NCHLT (Barnard et al., 2014), AfriSpeech (Olatunji et al., 2023), and NaijaVoices (Emezue et al., 2025) have expanded coverage across dozens of African languages, enabling the training and evaluation of multilingual speech models at previously impossible scales. Massively multilingual models such as MMS (Pratap et al., 2023) and wav2vec2 (Baevski et al., 2020) have demonstrated strong cross-lingual transfer, supporting hundreds of languages including many African ones through a single unified architecture. African-specific efforts such as AfriHUBERT (Alabi et al., 2024) and Simba (Elmadany et al., 2025) have further refined performance on African language subsets through targeted pretraining and fine-tuning. Despite these advances, evaluation of such models has focused predominantly on monolingual speech, leaving their behaviour on the code-switched speech that characterizes everyday African multilingual communication largely unexamined.

Code-Switching in Speech Code-switching — the alternation between two or more languages within a single conversation or utterance — is a defining feature of multilingual communities across Africa (Ndubuisi-Obi et al., 2019). Prior work on CS speech has produced datasets and benchmarks predominantly for South Asian language pairs such as Hindi-English (Dey and Fung, 2014) and Mandarin-English (Lyu et al., 2010), as well as European settings such as the Miami Bangor corpus (Deuchar et al., 2014). CS-FLEURS (Yan et al., 2025) extends FLEURS with code-switched speech across 100+ language pairs, providing the first large-scale multilingual CS speech benchmark. For African languages, CS speech resources remain scarce; notable exceptions include the South African soap opera CS corpus (van der Westhuizen and Niesler, 2018) and the Sepedi-English SPCS dataset (Modipa et al., 2015), both of which informed the explicit CS label design of Simba-49. CS-ASR has received growing attention (Yılmaz et al., 2018), with work demonstrating that models trained on monolingual data degrade substantially when applied to CS speech. CS-LID; identifying which languages are present in a CS utterance has received comparatively less attention, with most existing work focused on South Asian or European language pairs and few studies examining African language CS at scale (Burchell et al., 2024).

Multilabel and Synthetic Data Approaches Multilabel classification has been proposed for spoken language recognition by Nieto et al. (2023), who demonstrate that multilabel objectives improve efficiency for utterance-level LID on non-CS speech. Frame-level language diarization for CS speech has been explored for South African language pairs by Frost et al. (2022), requiring fine-grained temporal annotations. Synthetic CS data generation from monolingual sources has been explored for CS-ASR by Dhawan et al. (2023), who propose a concatenated tokenizer for token-level language identification during recognition. Our work extends the multilabel approach to African CS-LID specifically, combining synthetic CS training data with utterance-level multilabel fine-tuning and evaluating synthetic-to-real transfer on natural African-English CS speech.

To our knowledge, this work presents the first systematic evaluation of CS-LID for African language pairs and the first application of multilabel fine-tuning with synthetic CS data specifically for African spoken language identification.

3 Data

3.1 Monolingual evaluation corpus

We construct a monolingual evaluation corpus spanning 20 African languages drawn from publicly available speech datasets including AfriVox-1 (Awobade et al., 2026), Waxal (Diack et al., 2026), and Africa Next Voices (Nigeria, 2025; Marivate et al., 2026)[1] [2] [3], stratified by source, gender, and duration. We sample 1,000 utterances per language, yielding 20,000 utterances totalling 81.13

[1] https://huggingface.co/datasets/DigitalUmuganda/Afrivoice_Ethiopia
[2] https://huggingface.co/datasets/DigitalUmuganda/Afrivoice_Swahili
[3] https://huggingface.co/datasets/DigitalUmuganda/Afrivoice_Kinyarwanda

hours (Table 1). The corpus combines read speech and spontaneous speech drawn primarily from the Africa Next Voices and Waxal sources, providing coverage across diverse real-world recording conditions.

3.2 Natural CS evaluation corpus

For code-switched evaluation, we use a corpus of naturally occurring African-English speech collected from YouTube channels with high code-mixing frequency. Each utterance was validated through human review to ensure that it contains code-mixing between the matrix language and English. Each utterance is labeled with its matrix (dominant) language; no frame-level language-switch annotations are available. Dataset statistics can be found in (Table 2).

3.3 Synthetic CS training data

Due to the scarcity of natural CS speech for training, we construct synthetic code-switched audio by concatenating monolingual segments sourced from Waxal and AfricanNextVoices at controlled switch points. We generate two CS types: intersentential, involving a single switch point between two language segments, and multiswitch, involving three, five, or seven alternations within a single utterance. The multiswitch condition approximates intrasentential CS patterns more common in naturalistic African multilingual speech. Each condition is crossed with nine target durations (0.5–30s), three switch positions (10%, 50%, 90%), and three language ratios (10/90, 30/70, 50/50). For fine-tuning we use data from four African-English language pairs; Amharic-English, Hausa-English, Kinyarwanda-English, and Swahili-English; comprising 48,600 samples in total, split evenly between intersentential and multiswitch conditions (Table 3). The synthetic corpus is deliberately distinct from natural CS speech in prosody and discourse coherence; we treat synthetic-to-real transfer as an explicit evaluation condition rather than assuming in-domain training.

4 Models

We evaluate three spoken language identification models representing different approaches to African language coverage and architecture.

4.1 Simba-SLID-49

Simba-SLID-49 (Elmadany et al., 2025) is a spoken language identification model finetuned from AfriHUBERT (Alabi et al., 2024) on the SimbaBench multilingual training corpus (Elmadany et al., 2025), supporting 49 language labels. Uniquely among the models evaluated, Simba's label space includes five explicit code-switched labels covering South African language pairs (eng-zul, eng-xho, eng-sot, eng-tsn, nso-eng), trained on the CS Soap Opera (van der Westhuizen and Niesler, 2018) and SPCS (Modipa et al., 2015) datasets as part of its training corpus. This architectural distinction has direct implications for CS evaluation and is discussed in §6.1. Six languages in our evaluation set — English, Akan, French, Oromo, Nigerian Pidgin, and Shona — are absent from Simba's label space and are excluded from Simba's monolingual results.

Table 1: Composition of the monolingual evaluation set. Each language contains 1,000 utterances sampled from one or more public speech corpora.

Language	Family	Utt.	Src.	Hours
Afrikaans	Germanic	1000	4	1.80
Akan	Niger–Congo	1000	3	4.62
Amharic	Semitic	1000	5	3.62
Arabic	Semitic	1000	3	1.82
Fulani	Atlantic	1000	3	5.00
Hausa	Chadic	1000	6	9.52
Igbo	Volta–Niger	1000	5	7.60
Kinyarwanda	Bantu	1000	3	3.50
Luganda	Bantu	1000	3	3.49
Oromo	Cushitic	1000	2	4.90
Pedi	Bantu	1000	3	3.19
Pidgin	English Creole	1000	1	7.61
Sesotho	Bantu	1000	3	3.86
Shona	Bantu	1000	4	4.66
Swahili	Bantu	1000	5	3.41
Setswana	Bantu	1000	3	3.71
Wolof	Atlantic	1000	1	3.27
Xhosa	Bantu	1000	4	3.72
Yorùbá	Volta–Niger	1000	6	5.92
Zulu	Bantu	1000	4	3.91
Total	20	20,000	–	81.13

Language	Hours	# Samples
Amharic-English	6.2	745
Zulu-English	5.6	670
Kinyarwanda-English	3.8	464
Swahili-English	3.4	411
Hausa-English	2.2	273
Total	**21.2**	**2,563**

Table 2: Statistics of the natural code-switched evaluation dataset grouped by matrix language.

Table 3: Configuration of the synthetic code-switching training set.

Setting	Value
Language pairs	Amh–Eng, Hau–Eng, Kin–Eng, Swa–Eng
Durations (s)	0.5, 1, 2, 5, 10, 15, 20, 25, 30
Switch positions	10%, 50%, 90%
Language ratios	10/90, 30/70, 50/50
CS type	Intersentential, Multiswitch
Total samples	48,600

4.2 MMS-LID-256

MMS-LID-256 (Pratap et al., 2023) is a language identification model from Meta's Massively Multilingual Speech project, supporting 256 languages via a wav2vec2 encoder with a softmax classification head.

4.3 MMS-LID-1024

MMS-LID-1024 (Pratap et al., 2023) extends the MMS architecture to 1,024 languages.

4.4 Multilabel Fine-tuned MMS-256

To address the fundamental limitation of single-label prediction in CS settings, we fine-tune MMS-LID-256 with a multilabel objective. We replace the softmax classification head with a two-layer feedforward network (Linear(768, 768) → GELU → Dropout(0.1) → Linear(768, 256)) trained with binary cross-entropy loss (BCEWithLogitsLoss). The wav2vec2 feature extractor is frozen; only the top two transformer layers and the new classification head are updated. We train on synthetic CS data from four language pairs (Table 3) for 5 epochs using the Adam optimiser with a learning rate of 1×10^{-4} and a batch size of 16 on a single NVIDIA A100 GPU, completing in approximately 30 GPU hours. Language-weighted sampling compensates for corpus size imbalance: Kinyarwanda is upweighted at 3× (8,100 samples), Amharic at 1.5× (16,200 samples), and Swahili and Hausa at 1× (24,300 samples each). A 30% monolingual sample mixing rate per batch prevents catastrophic forgetting of monolingual identification capability.

4.5 Evalutation Metrics

- **Top-1 Accuracy**: Percentage of utterances where the model's top prediction matches the matrix language.
- **Both-in-Top-2**: Percentage of code-switched utterances where the matrix language is the top prediction and English appears in the model's top two predictions.
- **Both-in-Top-5**: Percentage of code-switched utterances where the matrix language is the top prediction and English appears in the model's top five predictions.

Table 4: Monolingual LID accuracy across 20 African languages. † indicates unsupported languages (accuracy set to 0). Bold denotes best performance per language.

Lang	Simba	MMS-256	MMS-1024
Afrikaans	**0.971**	0.961	0.948
Akan	†	0.844	**0.905**
Amharic	0.944	**0.980**	0.977
Arabic	0.669	0.969	**0.975**
Fulani	0.014	**0.929**	0.871
Hausa	0.828	0.987	**0.986**
Igbo	0.311	**0.946**	0.869
Kinyarwanda	0.986	0.992	**0.995**
Luganda	0.954	**0.996**	0.994
Oromo	†	0.940	**0.953**
Pedi	0.574	**0.818**	0.719
Pidgin	†	0.414	**0.532**
Sesotho	**0.619**	0.109	0.119
Shona	†	**0.983**	0.971
Swahili	0.894	**0.962**	0.957
Setswana	**0.650**	0.201	0.258
Wolof	0.448	0.805	**0.826**
Xhosa	0.630	**0.730**	0.590
Yoruba	0.387	**0.962**	0.945
Zulu	0.273	**0.685**	0.677
Mean (all)	0.508	**0.811**	0.803
Mean (sup.)	0.634	**0.811**	0.803

Table 5: Top-1 accuracy on the matrix (dominant) language for natural African–English code-switched speech. All models use single-label prediction; detection of the non-matrix language is near zero across all conditions.

Language	Simba-49	MMS-256	MMS-1024
Amharic	0.703	**0.992**	0.991
Hausa	0.744	**1.000**	**1.000**
Kinyarwanda	0.291	0.582	**0.679**
Swahili	0.847	**0.978**	0.966
Zulu	0.684	**0.745**	0.725
Mean	0.654	0.859	**0.872**

5 Results

We evaluate all models on monolingual LID, natural CS LID, reporting accuracy for monolingual and matrix identification tasks, and both-language accuracy for the multilabel finetuning evaluation.

5.1 Monolingual LID

Table 4 presents monolingual LID accuracy across 20 African languages. MMS-256 and MMS-1024 achieve comparable overall accuracy (0.811 and 0.803 respectively on all languages), while Simba-49 reaches 0.508 overall and 0.634 on its supported

Table 6: Top-1 matrix-language accuracy (Acc) and percentage of utterances where English appears in the top-5 predictions (Eng@5) for natural African–English code-switched speech. Hausa shows the most extreme suppression: English never appears in baseline top-5 predictions despite perfect matrix accuracy.

	MMS-256		MMS-1024	
Lang	**Acc**	**Eng@5**	**Acc**	**Eng@5**
Amh	**0.992**	0.038	**0.991**	0.019
Hau	**1.000**	0.000	**1.000**	0.000
Kin	0.582	0.075	**0.679**	0.132
Swa	**0.978**	0.080	0.966	0.090
Mean	0.888	0.048	**0.909**	0.060

subset. Performance varies substantially across languages: Afrikaans, Kinyarwanda, Amharic, and Luganda exceed 0.94 for all three models, while Zulu (0.273–0.677), Sesotho (0.109–0.619), and Setswana (0.201–0.650) show consistently lower accuracy. The low MMS accuracy on Sesotho and Tswana likely reflects label confusion with phonologically similar southern Bantu languages, a pattern consistent with findings in (Elmadany et al., 2025). These monolingual results establish the baseline against which CS performance degradation is measured.

5.2 Natural CS LID

Table 5 presents matrix language identification accuracy on natural African-English CS speech. MMS-256 and MMS-1024 achieve strong matrix accuracy on Amharic (0.992, 0.991) and Swahili (0.978, 0.966), but both models collapse on Kinyarwanda, dropping from near-perfect monolingual accuracy (0.992, 0.995) to 0.582 and 0.679 respectively; a degradation of up to 41 percentage points. Simba-49 shows more severe degradation across all languages, falling to 0.291 on Kinyarwanda and 0.847 on Swahili despite strong monolingual performance on both.

Across the MMS models and languages, English is identified in the top-5 predictions for fewer than 10% of utterances across most language pairs, indicating that models default to dominant-language detection and are largely blind to the switched component. We examine this behaviour in detail in §6.

Zulu results for Simba-49 require separate interpretation due to the presence of an explicit `eng-zul` label in Simba's label space; this is discussed in §6.1.

Table 7: Both-language detection before and after multilabel fine-tuning on natural African–English CS speech. Trained pairs were included in synthetic fine-tuning; Zulu is held-out. Both@5 measures whether both the matrix language (top-1) and English appear in the model's top-5 predictions. Δ = absolute improvement in Both@5.

	Matrix Acc		Both@5		Δ
Language	**Base**	**FT**	**Base**	**FT**	
Amharic	**.992**	.985	.038	**.997**	+.959
Hausa	**1.000**	.996	.000	**.996**	+.996
Kinyarwanda	.582	**.787**	.075	**.994**	+.919
Swahili	**.978**	.893	.080	**.888**	+.808
Overall	.888	**.915**	.048	**.969**	+.921
Zulu†	**.745**	.730	.018	**.357**	+.339

† Held-out transfer case; not seen during fine-tuning.

5.3 Multilabel Fine-tuning

Prior to fine-tuning, MMS-256 identifies both the matrix language and English in its top-5 predictions for only 4.8% of natural CS utterances across the four evaluation pairs, confirming that the single-label model has no meaningful sensitivity to the non-matrix language regardless of its monolingual performance.

Trained pairs. After multilabel fine-tuning on synthetic CS data, both-in-top-5 rises to 96.9% overall across the four trained pairs. a gain of 92.1 percentage points with Amharic reaching 99.7%, Hausa 99.6%, Kinyarwanda 99.4%, and Swahili 88.8% (Table 7). Hausa is the starkest case: English appears in zero baseline top-5 predictions across 273 utterances, rising to 99.6% after fine-tuning with matrix accuracy fully preserved at 99.6%. Kinyarwanda shows the most dramatic recovery across both metrics: matrix accuracy improves from 0.582 to 0.787, and both-in-top-5 rises from 0.075 to 0.994, demonstrating that fine-tuning on synthetic CS data resolves the substantial CS-induced degradation observed in the single-label baseline.

Swahili matrix accuracy shows a real and expected trade-off, declining from 0.978 to 0.893 (-0.085). This is a direct consequence of the multilabel objective: a model that assigns probability mass to both code-switched languages will necessarily be less concentrated on the dominant language than a single-label model. This cost is the price of bilingual sensitivity and should be interpreted as such, not as degradation.

Held-out transfer: Zulu-English. Zulu-English was withheld entirely from synthetic fine-tuning to assess zero-shot transfer. Both-in-top-5 rises from 1.8% to 35.7% a gain of 33.9 percentage points, while matrix accuracy is largely preserved (0.745 to 0.730). While this falls well short of trained-pair performance, it represents meaningful generalisation: the model acquires bilingual sensitivity to an unseen language pair from exposure to four typologically diverse training pairs spanning Semitic (Amharic), Chadic (Hausa), and Bantu (Kinyarwanda, Swahili) families. The factors limiting Zulu transfer are examined in §6.3.

Monolingual preservation. Monolingual identification capability is fully preserved throughout training. Top-1 accuracy on held-out monolingual speech remains at 0.997, 0.996, 0.963, and 0.992 for Amharic, Kinyarwanda, Swahili, and Hausa respectively matching or exceeding the original MMS-256 baseline confirming that the 30% monolingual batch mixing strategy prevents catastrophic forgetting.

6 Analysis

6.1 Two Paradigms, Two Failure Modes

Analysis of model predictions reveals two distinct failure modes that motivate the multilabel fine-tuning approach.

Failure Mode 1: Explicit label generalisation failure. Exemplified by Simba-49. On Zulu-English CS speech, 68.4% of Simba's predictions are the explicit label `eng-zul`, trained on South African soap opera data (van der Westhuizen and Niesler, 2018). The model identifies CS correctly on this pair, but only because it was trained on it directly. This approach does not generalise: Simba has no corresponding label for Amharic-English, Kinyarwanda-English, or Swahili-English. Explicit CS labels require paired CS training data for every language combination, which is precisely the resource scarcity problem the field faces.

Failure Mode 2: Softmax suppression failure. Exemplified by MMS-256 and MMS-1024. On Zulu-English CS speech, MMS-1024's top-5 predictions are drawn exclusively from southern Bantu languages (`zul`, `xho`, `nbl`, `ssw`, `nso`). English present in every utterance appears in fewer than 3% of top-5 predictions. The model is not uncertain between the two CS languages; it does not register English as present at all. This is a direct consequence of the softmax objective, which forces zero-sum competition between 1,024 classes. When the dominant language wins the competition, it suppresses all other signals including a language that occupies a substantial portion of the utterance.

These two failure modes represent opposite ends of a design spectrum: one model knows about CS but only for specific trained pairs, while the other has broad language coverage but cannot represent bilingual speech by architectural design.

6.2 Fine-tuning Resolution: The Hausa Case

To illustrate the qualitative change in model behaviour after fine-tuning, we examine top-5 prediction distributions for Hausa-English CS speech. Table 8 shows the dominant prediction patterns before and after fine-tuning across all 273 Hausa-English utterances.

In the baseline model, rank 1 is `hau` in 100% of utterances and rank 2 is `yor` in 99% of utterances. Ranks 3–5 are occupied entirely by unrelated languages: Zarma (`dje`), Pashto (`pus`), Shona (`sna`), Urdu (`urd`), and Mandarin (`cmn`). English does not appear in a single top-5 prediction across all 273 utterances. These filler predictions are not random noise, they reflect the softmax suppression failure mode directly: once Hausa dominates rank 1, probability mass concentrates on acoustically proximate or co-occurring languages rather than distributing toward English, which is genuinely present in every utterance.

After fine-tuning, the prediction space is restructured entirely. The dominant pattern is `hau|amh|eng|yor|kin`, accounting for the majority of utterances, with English present in the top-5 for 99.6% of predictions. Ranks 3–5 are now occupied by languages from the fine-tuning training set and geographically proximate languages. Hausa matrix accuracy is preserved at 99.6%. This before/after comparison demonstrates that multilabel fine-tuning does not merely append English to the tail of the softmax distribution: it restructures the model's probability space around the languages genuinely present in the utterance.

6.3 Partial Transfer to Unseen Pairs: Zulu-English

Zulu-English was withheld entirely from synthetic fine-tuning to assess zero-shot transfer. Both-in-top-5 rises from 1.8% to 35.7%; a gain of 33.9 percentage points, while matrix accuracy is largely

Table 8: Dominant top-5 prediction patterns for Hausa-English CS utterances before and after multilabel fine-tuning. Baseline ranks 3–5 are occupied by unrelated languages; fine-tuning replaces these with languages present in or proximate to the training set.

Rank	Baseline	Fine-tuned
1	hau (100%)	hau (99.6%)
2	yor (99%)	amh or eng
3–5	dje, pus, sna	eng, yor, kin
English present	0%	99.6%

preserved (0.745 to 0.730). While this falls substantially short of trained-pair performance, it represents meaningful generalisation: the model acquires bilingual sensitivity to an unseen language pair from exposure to four typologically diverse training pairs.

We attribute the transfer gap to two factors. First, Zulu belongs to the southern Bantu subgroup, which is phonologically and lexically distinct from the Bantu training pairs (Kinyarwanda, Swahili); transfer across subgroups is more demanding than within-subgroup transfer. Second, MMS-256 monolingual Zulu accuracy (0.685) is the lowest among all evaluated languages, indicating that Zulu occupies a less well-defined region of the model's representation space, a weaker foundation for any fine-tuning signal to build on. Extending synthetic fine-tuning to include Zulu monolingual audio, which is available in our corpus, is expected to recover performance closer to trained-pair levels and is left to future work.

6.4 Non-matrix Language Suppression

Table 6 quantifies the extent to which non-matrix language evidence remains marginal in baseline model predictions. Across all models and evaluated language pairs, English appears in the top-5 predictions for only a small fraction of natural CS utterances despite being present in every clip, reflecting strong concentration of probability mass around the matrix language and related language clusters.

The relationship between matrix accuracy and English top-5 presence is suggestive but not uniform. Among MMS-256 predictions, languages with higher matrix accuracy; Amharic (0.992) and Swahili (0.978); show lower English top-5 presence (0.038 and 0.080), consistent with dominant-language confidence suppressing non-matrix signals. Kinyarwanda shows both low matrix accuracy (0.582) and low English presence (0.075), indicating that factors beyond matrix-language confidence, such as acoustic distance between the matrix language and English, also govern non-matrix signal visibility. The Hausa case is the most extreme: despite perfect matrix accuracy (1.000), English never appears in the baseline top-5 across 273 utterances, confirming that strong monolingual discrimination actively suppresses bilingual signal rather than leaving space for it.

The deployment implication is direct: improvements in monolingual language discrimination do not improve and may actively reduce sensitivity to CS speech. Increasingly strong matrix-language representations risk causing bilingual utterances to be interpreted as progressively more monolingual.

7 Conclusion

We present a systematic evaluation of spoken language identification on monolingual and code-switched African speech, establishing a benchmark across 20 languages and five natural African-English CS pairs. Two fundamental failure modes characterise existing approaches: explicit CS label models achieve strong performance on trained pairs but do not generalise, while monolingual label space models suppress non-matrix language signals entirely, with both-in-top-5 accuracy below 10% across all evaluated pairs for baseline models.

Multilabel fine-tuning on synthetic CS data addresses both failure modes without requiring natural CS audio for training. Across four African-English pairs, both-in-top-5 rises from 4.8% to 96.9%, with monolingual identification capability fully preserved. Partial zero-shot transfer to a held-out pair (Zulu-English: 35.7% vs. 1.8% baseline) demonstrates that the approach acquires generalisable bilingual sensitivity beyond trained language pairs.

Future work should extend evaluation to African-African CS pairs, investigate frame-level language detection with annotated switch points and assess the multilabel approach on additional spoken LID architectures.

Limitations

This work has several limitations. First, multilabel fine-tuning is evaluated on four African-English language pairs; generalisation to African-African CS pairs remains undemonstrated on natural data, and findings may not extend to settings where

both languages are African. Second, synthetic CS training data consists of concatenated monolingual segments lacking the prosodic continuity and discourse coherence of naturalistic CS speech; while synthetic-to-real transfer is demonstrated, the ceiling imposed by this domain gap is unknown. Third, both-in-top-5 measures English presence anywhere in the top-5 predictions rather than confident bilingual detection. English sigmoid scores are consistently lower than matrix language scores, indicating partial rather than fully symmetric bilingual sensitivity. Fourth, fine-tuning is performed on MMS-LID-256 only; whether the multilabel approach generalises to other spoken LID architectures requires further investigation.

Ethics Statement

Potential risks. The primary intended use of this work is to improve spoken language identification in African multilingual settings, enabling more reliable downstream speech systems. A potential misuse risk is that more accurate LID could facilitate surveillance of multilingual speakers; we encourage deployers to consider privacy implications and to obtain appropriate consent before applying LID systems to personal or intercepted speech.

Use of existing artifacts. All datasets used in this work; AfriVox-1, Waxal, Africa Next Voices, and AfriSpeech, are publicly released research datasets accessed under their published licences (Creative Commons or equivalent open research licences), consistent with their stated research purpose. The synthetic CS corpus derived from these sources is likewise intended solely for research use and inherits the access conditions of its source materials; it should not be used outside research contexts.

Personal and sensitive data. The monolingual and code-switched evaluation corpora consist of speech recordings released by their respective dataset creators; we did not collect new data involving human participants. The YouTube-sourced CS evaluation corpus was collected from publicly posted videos; human reviewers validated language content at the utterance level without retaining speaker metadata. No steps were taken to re-identify speakers, and no sensitive personal information was transcribed or stored.

Annotation instructions and participant consent. We did not recruit or pay external participants or annotators. All speech data originates from existing datasets whose collection procedures and consent processes are described in their respective publications. The CS evaluation corpus was validated by internal team members, who were asked to confirm the presence of code-mixing between the stated matrix language and English within each utterance; no crowdsourcing platform was used.

Descriptive statistics. All accuracy figures reported in Tables 4; 7 are means computed over the full evaluation sets described in Section 3, with sample sizes given in Tables 1 and 2. Results reflect a single training run; no cross-seed variance is reported, and all figures should be interpreted as point estimates.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages. *Preprint*, arXiv:2409.20201.

Rosana Ardila and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*.

Busayo Awobade, Mardhiyah Sanni, Tassallah Abdullahi, Chibuzor Okocha, Kelechi Ezema, Devendra Deepak Kayande, Lukman Enegi Ismaila, Tobi Olatunji, and Gloria Ashiya Katuka. 2026. AfriVox: Probing multilingual and accent robustness of speech LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2672–2690, Rabat, Morocco. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The NCHLT speech corpus of the South African languages. In *4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 194–200.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian's, Malta. Association for Computational Linguistics.

Alexis Conneau and 1 others. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *ICASSP*.

Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and Diana Carter. 2014. *5. Building Bilingual Corpora*, pages 93–110. Multilingual Matters, Bristol, Blue Ridge Summit.

Anik Dey and Pascale Fung. 2014. A Hindi-English code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kunal Dhawan, KDimating Rekesh, and Boris Ginsburg. 2023. Unified model for code-switching speech recognition and language identification based on concatenated tokenizer. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 74–82, Singapore. Association for Computational Linguistics.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, and 24 others. 2026. Waxal: A large-scale multilingual african language speech corpus. *Preprint*, arXiv:2602.02734.

AbdelRahim A. Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Mapping Africa's speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11039–11061, Suzhou, China. Association for Computational Linguistics.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Toluwase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *CoRR*, abs/2505.20564.

Geoffrey Frost, Emily Morris, Joshua Jansen van Vüren, and Thomas Niesler. 2022. *Fine-Tuned Self-supervised Speech Representations for Language Diarization in Multilingual Code-Switched Speech*, page 246–259. Springer Nature Switzerland.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in code-switching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 87–93, Doha, Qatar. Association for Computational Linguistics.

Shuyue Stella Li, Cihan Xiao, Tianjian Li, and Bismarck Odoom. 2023. Simple yet effective code-switching language identification with multitask pre-training and transfer learning. *Preprint*, arXiv:2305.19759.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. 2014. The CMU submission for the shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 80–86, Doha, Qatar. Association for Computational Linguistics.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. SEAME: a Mandarin-English code-switching speech corpus in south-east asia. In *Interspeech 2010*, pages 1986–1989.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2005–2011, Minneapolis, Minnesota. Association for Computational Linguistics.

Vukosi Marivate, Kayode Olaleye, Sitwala Mundia, Andinda Bakainga, Unarine Netshifhefhe, Mahmooda Milanzie, Tsholofelo Hope Mogale, Thapelo Sindane, Zainab Abdulrasaq, Kesego Mokgosi, Chijioke Okorie, Nia Zion Van Wyk, Graham Morrissey, Dale Dunbar, Francois Smit, Tsosheletso Chidi, Rooweither Mabuya, Andiswa Bukula, Respect Mlambo, and 4 others. 2026. Swivuriso: The south african next voices multilingual speech dataset. *Preprint*, arXiv:2512.02201.

T. I. Modipa, M. H. Davel, and F. De Wet. 2015. SPCS speech corpus. SADiLaR Language Resource Repository, License: Creative Commons Attribution 2.5 South Africa license: https://creativecommons.org/licenses/by/2.5/legalcode.

Innocent Ndubuisi-Obi, Sayan Ghosh, and David Jurgens. 2019. Wetin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior in Nigerian online discussions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6204–6214, Florence, Italy. Association for Computational Linguistics.

Oriol Nieto, Zeyu Jin, Franck Dernoncourt, and Justin Salamon. 2023. Efficient spoken language recognition via multilabel classification. *arXiv preprint arXiv:2306.01945*.

DataScience Nigeria. 2025. African voices: Multilingual speech dataset for low-resource african languages. `https://www.africanvoices.ai`. Latest release, November 2025.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Ewald van der Westhuizen and Thomas Niesler. 2018. A first South African corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, Shinji Watanabe, Chih-Chen Chen, and 8 others. 2025. Cs-fleurs: A massively multilingual and code-switched speech dataset. *Preprint*, arXiv:2509.14161.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a Unified Code-Switching ASR System for South African Languages. In *Interspeech 2018*, pages 1923–1927.